Complémentarité des méthodes numériques et symboliques en pharmacovigilance

Jean Villerd[1], Yannick Toussaint[1], Agnès Lillo-Le Louët[2]

[1] Loria - INRIA Nancy Grand Est, Nancy
{nom.prenom}@loria.fr
[2] CRPV, Hôpital Européen Georges Pompidou, Paris
agnes.lillo-loelouet@egp.aphp.fr

Résumé : « La pharmacovigilance a pour objet la surveillance du risque d'effet indésirable résultant de l'utilisation des médicaments et produits à usage humain[1] ». Elle consiste à détecter des « signaux » correspondant à des corrélations suspectes entre la prise d'un médicament et l'observation d'un effet indésirable. Les professionnels de santé sont tenus de signaler tout effet indésirable inattendu en envoyant un rapport à leur centre régional de pharmacovigilance, alimentant ainsi une base nationale. Pour faire face à la quantité de rapports stockés, des méthodes numériques fondées sur des mesures de disproportionnalité ont été proposées ; elles extraient efficacement les couples (médicament, effet indésirable) sur-représentés dans la base, en vue de leur évaluation par un expert. Toutefois, ces approches fournissent peu d'information aux experts pour leur permettre de comprendre pour quelles raisons ces signaux ont été retenus. De plus, aucune information ne leur est transmise concernant la spécificité éventuelle d'un signal qui concernerait plus particulièrement une sous-population (sexe, âge).

L'utilisation d'une approche symbolique comme l'Analyse de Concepts Formels (ACF), en complément des méthodes numériques existantes, nous a permis d'obtenir une meilleure « traçabilité » des signaux extraits en fournissant aux experts une vue synthétique du contenu de la base. De plus, l'utilisation d'une approche fondée sur les treillis permet de réduire l'espace de recherche, de détecter des associations complexes impliquant plusieurs médicaments et effets indésirables et d'identifier les spécificités liées aux diverses sous-populations. Nous illustrons ces atouts par une expérimentation sur la base régionale HEGP.

Mots-clés : Extraction de connaissances, Pharmacovigilance, Fouille de données

1 Introduction

Depuis une dizaine d'années, des méthodes de fouille de données sont appliquées aux bases de pharmacovigilance. Dans cet article, nous montrons que l'introduction d'une méthode symbolique telle que l'Analyse de Concepts Formels (ACF), en complément

[1] Article R. 5121-150 du code de la santé publique

TAB. 1 – Table de contingence du couple (d, e)

	e	$\overline{e}$	
d	x	y	$x+y$
$\overline{d}$	z	t	$z+t$
	$x+z$	$y+t$	N

des approches numériques existantes, améliore l'efficacité de celles-ci et fournit des informations supplémentaires facilitant l'évaluation des résultats par les experts.

La pharmacovigilance veille à la sécurité d'emploi des médicaments et consiste à (i) collecter et gérer les données relatives à l'emploi des médicaments (ii) exploiter ces données afin de détecter tout fait nouveau concernant la sécurité des médicaments. Une des problématiques en pharmacovigilance est la détection des effets indésirables des médicaments, également appelée « détection du signal ». Il s'agit d'identifier des « signaux », i.e. des couples (d, e) pour lesquels on observe une corrélation entre l'apparition d'un effet indésirable e et la prise d'un médicament d. Les données sont constituées de rapports de signalement. Un rapport contient des informations sur le profil du patient, les médicaments prescrits, et l'effet indésirable observé. Compte tenu de la quantité de rapports stockés dans la base nationale, l'exploitation manuelle de celle-ci est extrêmement coûteuse. En effet, en 2008 plus de 20 000 nouveaux rapports sont venus alimenter la base française, gérée par l'Afssaps ; la base de l'OMS contient elle plus de 3 millions de rapports.

Face à cette situation, des méthodes automatiques de fouille de données ont été développées, extrayant automatiquement des signaux potentiels devant être évalués par des experts (Hauben *et al.*, 2005; Bate *et al.*, 2008). Ces méthodes sont fondées sur des mesures statistiques largement utilisées en pharmacovigilance (*Proportional Reporting Ratio*, *Risk Odds Ratio*, χ^2) et sur des approches numériques telles que les réseaux bayésiens. Bien que leur efficacité ait été prouvée, elles présentent encore quelques lacunes. L'impact des aspects démographiques est sous-exploité, le filtrage des signaux dûs au bruit reste à améliorer, et leur fonctionnement en « boîte noire » ne facilite pas l'évaluation des résultats par les experts. En outre, peu de travaux se sont penchés sur les signaux d'ordres supérieurs impliquant plusieurs médicaments et effets indésirables tels que les interactions médicamenteuses.

L'utilisation d'une approche symbolique comme l'AFC, en complément des méthodes numériques existantes, permet d'obtenir une meilleure « traçabilité » des signaux extraits en fournissant aux experts une représentation synthétique du contenu de la base. En effet, d'un point de vue symbolique, la détection du signal peut être appréhendée comme un problème d'extraction de motifs. Les motifs pertinents étant ceux qui satisfont un critère de disproportionnalité. Nous montrons que l'utilisation d'une approche fondée sur les treillis permet en outre de réduire l'espace de recherche, de comparer les motifs selon leur granularité (nombre de médicaments, nombre d'effets), et de caractériser leur éventuelle spécificité à une sous-population particulière.

2 État de l'art

Cette section définit le vocabulaire utilisé en pharmacovigilance, présente les méthodes de détection automatique du signal existantes et montre leurs points faibles.

2.1 Vocabulaire

Bien que la notion de « signal » joue un rôle central en pharmacovigilance, ce terme souffre d'un manque de définition consensuelle. Une définition précise est avancée par Meyboom *et al.* (1997) : « un ensemble de données générant une hypothèse significative au regard de l'utilisation rationnelle d'un médicament ». Suivant cette définition, un **signal** se compose (i) d'un couple (d, e) où le médicament d est suspecté d'être la cause de l'effet indésirable e (hypothèse), (ii) d'un ensemble de rapports (données) et (iii) de mesures statistiques (arguments). Dans la suite, nous considérerons également des hypothèses d'interactions médicamenteuses de la forme $(d_1 d_2, e)$.

Une **signal (resp. interaction) potentiel** est un signal (resp. interaction) satisfaisant un critère de pertinence donné et qui sera présenté aux experts pour évaluation. À notre connaissance, les méthodes existantes traitent uniquement soit des signaux, soit des interactions. La sous-section suivante présente les critères de pertinence utilisés par les méthodes de détection automatique du signal existantes.

2.2 Méthodes et mesures

L'objectif des méthodes de détection automatique du signal est d'identifier, parmi tous les couples (d, e), ceux dont le nombre d'occurrences dans la base est supérieur au nombre d'occurrences attendu si d et e étaient indépendants. Cependant, si le nombre d'occurrences de (d, e) dans la base est connu, ni le nombre total de patients ayant pris d ni le nombre de patients souffrant de e dans la population totale n'est connu. En effet la base ne contient un rapport mentionnant d uniquement si un effet indésirable a été observé chez un patient prenant d. Dès lors, le nombre attendu d'occurrences de (d, e) ne peut être calculé de façon précise (Roux *et al.*, 2005). Une solution consiste à estimer le nombre d'occurrences attendu de (d, e) en considérant le nombre de rapports concernant d'autres médicaments et d'autres effets indésirables présents dans la base. Les tables de contingence jouent ainsi un grand rôle en pharmacovigilance. La Table 1 décrit la table de contingence du couple (d, e). Chaque cellule contient le nombre de rapports contenant la combinaison correspondant. Ainsi, x est le nombre de rapports contenant à la fois d et e, y est le nombre de rapports contenant d mais pas e, etc. N est le nombre total de rapports. Les méthodes existantes diffèrent dans le choix du critère statistique utilisé pour quantifier la surreprésentation d'un couple. Le plus employé est une mesure de disproportionnalité appelée *Proportional Reporting Ratio (PRR)* (Evans *et al.*, 2001), défini par $PRR(d, e) = \frac{P(e|d)}{P(e|\overline{d})} = \frac{\frac{x}{x+y}}{\frac{z}{z+t}}$. Le couple (d, e) est considéré comme un signal potentiel si : $PRR \geq 3$ et $\chi^2 \geq 4$ et $x \geq 3$ (Evans *et al.*, 2001; Hauben *et al.*, 2005). Intuitivement, la première condition signifie que pour que (d, e) soit considéré comme un signal potentiel, il doit y avoir deux fois plus de chances de souffrir de e sachant qu'on a pris d que de chances de souffrir de e sachant qu'on n'a pas pris

d. La deuxième correspond au test d'indépendance du χ^2. La troisième considère que (d, e) doit apparaître au moins trois fois dans la base. Des méthodes plus sophistiquées implémentent les mesures de disproportionnalité dans un cadre bayésien (DuMouchel, 1999). Cependant, toutes présentent les lacunes décrites dans la section suivante.

3 Limites des méthodes existantes

Cette section traite des limites des méthodes existantes, notamment du nombre important de couples testés ainsi que de la difficulté à prendre en compte les aspects démographiques. Enfin nous montrons que, pour évaluer efficacement les résultats, les experts ont besoin d'informations complémentaires à la liste des mesures statistiques.

Espace de recherche

Les méthodes existantes testent tous les couples possibles $(d, e) \in \mathcal{D} \times \mathcal{E}$, où $\mathcal{D}$ est l'ensemble des médicaments et $\mathcal{E}$ l'ensemble des effets indésirables qui apparaissent dans la base. Ainsi le nombre de couples à évaluer est $|\mathcal{D}| \cdot |\mathcal{E}|$. La même stratégie appliquée aux associations d'ordres supérieurs (D, E), où $D \subseteq \mathcal{D}$ et $E \subseteq \mathcal{E}$ sont des ensembles non vides, conduit à évaluer $(2^{|\mathcal{D}|} - 1) \cdot (2^{|\mathcal{E}|} - 1)$ associations candidates, ce qui peut être extrêmement coûteux sur de grandes bases.

Signaux et interactions

Peu de travaux se sont penchés sur les associations autres que les signaux. Toutefois, considérer une interaction $(d_1 d_2, e)$ soulève les questions suivantes au sujet de ses signaux associés (d_1, e) et (d_2, e) :
(i) dans quelle mesure est-il pertinent d'évaluer le signal (d_1, e) si tous les patients qui ont pris d_1 et souffrent de e ont aussi pris d_2 ? (ii) dans quelle mesure est-il pertinent d'évaluer le signal (d_1, e) si tous les patients qui ont pris d_1 et souffrent de e ont aussi pris d'autres médicaments différents d_2, d_3 ?

Aspects démographiques

Les attributs démographiques tels que le sexe ou l'âge peuvent avoir un impact particulier. Une association (D, E, W) peut être spécifique à la sous-population W. Des travaux ont montré la nécessité d'examiner les signaux sur des sous-populations différentes, appelées *strates*, mais se sont heurtés aux problèmes combinatoires liés à un espace de recherche exhaustif. L'évaluation de (d, e, W) sur chaque strate différente est non seulement coûteuse mais de plus il n'est pas réaliste de considérer que chaque strate doive être évaluée individuellement par un expert (Bate *et al.*, 2008).

Évaluation

L'évaluation est la tâche à l'issue de laquelle les experts décident si un signal potentiel nécessite une expertise approfondie par des essais cliniques. Il s'agit d'une tâche difficile et coûteuse en temps. Selon la définition de Meyboom, un signal se compose

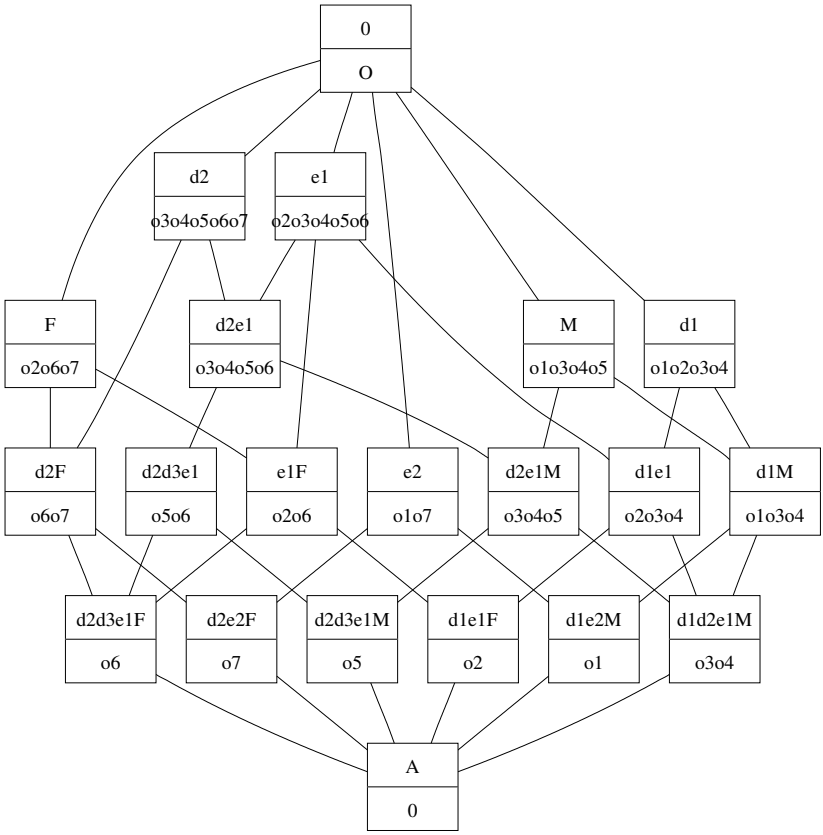

FIG. 1 – Treillis de concepts associé au contexte du tableau 2

d'un hypothèse étayée par des données et des arguments. Malheureusement, les seuls arguments fournis à l'expert par les méthodes numériques existantes sont les valeurs des mesures de disproportionnalité.

Les experts doivent bénéficier d'un dispositif leur permettant de remettre le signal potentiel dans son contexte d'extraction, i.e. d'assurer la traçabilité d'un signal potentiel extrait à partir de la base de rapports. Il doivent disposer d'une représentation synthétique de la base leur permettant d'appréhender les conditions dans lesquelles un signal potentiel a été identifié.

Enfin, la base peut être bruitée. En effet, les rapports sont à l'origine manuscrits et peuvent être mal renseignés.

4 Analyse de concepts formels (AFC)

Nous présentons ici l'Analyse de concepts formels (AFC), une famille de méthodes de classification symbolique que nous introduisons en complément des mesures de dis-

proportionnalité. Après les définitions des notions de base, nous montrons comment elle peut être utilisée dans le cadre de la pharmacovigilance.

4.1 Définitions

Considérant une relation binaire entre un ensemble d'objets $\mathcal{O}$ et un ensemble d'attributs $\mathcal{A}$, l'ACF extrait un ensemble de couples (O, A) avec $O \subseteq \mathcal{O}$, $A \subseteq \mathcal{A}$, appelés concept formels, tels que chaque objet O possède tous les attributs de A et vice-versa. Les concepts formels, partiellement ordonnés selon l'inclusion de O et A, forment une structure de treillis appelée treillis de concepts. Ce treillis peut être vu comme une conceptualisation de la relation binaire.

Nous reprenons les définitions de Ganter & Wille (1999). Un **contexte formel** est un triplet $\mathcal{K} = (\mathcal{O}, \mathcal{A}, I)$ où $\mathcal{O}$ est un ensemble d'**objets**, $\mathcal{A}$ un ensemble d'**attributs**, et $I \subseteq \mathcal{O} \times \mathcal{A}$ une relation binaire telle que oIa si l'objet o possède l'attribut a.
Le Tableau 2 représente un contexte formel $\mathcal{K}$ avec $\mathcal{O} = \{o_1 \dots o_7\}$ et $\mathcal{A} = \{d_1 \dots d_3\} \cup \{e_1, e_2\} \cup \{M, F\}$.

Deux **opérateurs de dérivation**, tous deux notés $(.)'$, lient objets et attributs. Considérant un ensemble d'objets $O \subseteq \mathcal{O}$, $O' = \{a \in \mathcal{A} | oIa\}$, i.e. O' est l'ensemble d'attributs partagés par tous les objets de O. De manière duale, $A' = \{o \in \mathcal{O} | oIa\}$ est l'ensemble d'objets possèdant tous les attributs de A.Par exemple, $\{d_1, d_2\}' = \{o_3, o_4\}$ et $\{o_3, o_4\}' = \{d_1, d_2, e_1, M\}$.

Un ensemble d'attributs A est dit **fermé** si $A = A''$. L'ensemble des ensembles B tels que $B'' = A$ forme la **classe d'équivalence** de A. Tous les ensembles de la classe d'équivalence de A ont le même **support** $\sigma(A)$ que A, i.e. la cardinalité de A'. Par exemple, $\{d_1, d_2\}$ n'est pas fermé puisque $\{d_1, d_2\}'' = \{o_3, o_4\}' = \{d_1, d_1, e_1, M\}$, tandis que $\{o_3, o_4\}$ est fermé puisque $\{o_3, o_4\}'' = \{d_1, d_2, e_1, M\}' = \{o_3, o_4\}$.

Un **concept formel** est un couple (O, A) tel que $O = O''$ et $A = O'$. Chaque objet de O tous les attributs de A et vice-versa. Notons que $A'' = (O')'' = (O'')' = O' = A$, ainsi O et A sont fermés. O (resp. A) est appelé l'**extension**, notée $\text{Ext}(O, A)$ (resp. l'**intension** notée $\text{Int}(O, A)$ du concept. L'ensemble des concepts formels du contexte formel $\mathcal{K}$ est noté $\mathfrak{B}(\mathcal{K})$. Par exemple, $(\{o_3, o_4\}, \{d_1, d_2, e_1, M\})$ est un concept formel.

Les concepts formels sont partiellement ordonnés selon l'inclusion de leurs extensions. Considérant deux concepts (O_1, A_1) et (O_2, A_2), $(O_1, A_1) \leq (O_2, A_2)$ ssi $O_1 \subseteq O_2$ (ou de manière équivalente $A_1 \supseteq A_2$). L'ensemble des concepts ainsi ordonnés est noté $\underline{\mathfrak{B}}(\mathcal{K})$ et est appelé le **treillis de concepts** du contexte formel $\mathcal{K}$. Le concept maximal $(\mathcal{O}, \mathcal{O}')$ est appelé le concept *top*, et le concept minimal $(\mathcal{A}', \mathcal{A})$ est appelé le concept *bottom*.

Le treillis de concepts $\underline{\mathfrak{B}}(\mathcal{K})$, associé au contexte $\mathcal{K}$ est illustré par la Figure 1. Chaque nœud représente un concept formel avec son intension en partie haute et son extension en partie basse.

Considérant un attribut a, son **concept attribut**, noté $\mu(a)$, est l'unique concept (a'', a'), i.e. sur la Figure 1 le concept le plus haut contenant a en intension. Par exemple, $\mu(e_2) = (\{o_1, o_7\}, \{e_2\})$.

TAB. 2 – Exemple de contexte formel

	d_1	d_2	d_3	e_1	e_2	M	F
o_1	×				×	×	
o_2	×			×			×
o_3	×	×		×		×	
o_4	×	×		×			×
o_5		×	×	×		×	
o_6		×	×	×			×
o_7		×			×		×

Dans le pire des cas, le nombre de concepts associés à $\mathcal{K} = (\mathcal{O}, \mathcal{A}, I)$ est $2^{\min(|\mathcal{O}|,|\mathcal{A}|)}$, lorsque chaque sous-ensemble de $\mathcal{O}$ ou $\mathcal{A}$ est fermé, ce qui est improbable en pratique.

4.2 Détection du signal et fouille de données

La détection du signal peut être appréhendée comme un problème de fouille de données, consistant à identifier des **motifs** (i.e. des ensembles d'attributs) pertinents parmi un ensemble de rapports. Les motifs pertinents doivent contenir au moins un médicament et un effet indésirable, et satisfaire un critère de pertinence correspondant à des seuils sur des mesures de disproportionnalité.

Dans ce cadre, l'AFC présente des propriétés intéressantes. Tout d'abord le treillis définit un espace de recherche réduit : examiner uniquement les motifs fermés est suffisant pour identifier tous les motifs pertinents. Ensuite l'ordre partiel entre motifs fermés fournit toutes les données nécessaires pour construire les tables de contingence et calculer les mesures de disproportionnalité. Enfin, l'ordre partiel fournit les liens nécessaires à la comparaison de motifs de longueurs différentes, comme les signaux et interactions.

5 Extraction des signaux et interactions à partir du treillis

Notre méthode extrait les signaux et interactions potentiels à partir du treillis de concepts correspondant au contexte $(\mathcal{O}, \mathcal{A}, I)$, où $\mathcal{O}$ est l'ensemble des rapports et $\mathcal{A} = \mathcal{D} \cup \mathcal{E} \cup \mathcal{W}$ l'ensemble des attributs. Nous justifions dans un premier temps la réduction de l'espace de recherche aux seuls motifs fermés, puis définissons des règles de préférence entre signaux et interactions.

5.1 Restriction aux motifs fermés

De par la construction du treillis, l'espace de recherche des signaux est restreint aux motifs fermés, i.e. aux intensions des concepts formels. Nous montrons ci-dessous que deux situations peuvent se produire. Dans les deux cas, la restriction aux motifs fermés n'engendre pas de perte d'information : ni perte de signal, ni perte d'interaction.

– Dans le premier cas, considérons que le motif fermé contient un attribut démographique supplémentaire qui fait que l'association concerne une sous-population plus spécifique

que celle du non-fermé. Cela signifie que le non-fermé étudié n'apparait jamais dans une population plus grande. Considérons, par exemple, l'interaction fermée $(d_1 d_2 e_1 M)$ (voir Figure 1). Sa classe d'équivalence contient aussi les motifs $(d_1 d_2 e_1)$ et $(d_1 e_1 M)$. Par définition de la fermeture, ces trois motifs sont partagés par le même ensemble de rapports $\{o_3, o_4\}$ et ont donc le même support. Cela signifie que $(d_1 d_2 e_1)$ n'a été observée que sur des hommes et l'étudier sur l'ensemble de la population n'est donc pas pertinent. En revanche, $(d_1 d_2 e_1 M)$ donne une description très exacte de la sous-population sur laquelle cette interaction a été effectivement observée.

– Dans le second cas, le motif fermé porte sur la même population que le fermé (pas d'attribut démographique supplémentaire) et le non-fermé est alors moins pertinent (son PRR est inférieur au PRR de son fermé). Considérons le cas du motif $(d_1 e_1 M)$ qui serait un signal non fermé et dont la fermeture est $(d_1 d_2 e_1 M)$. Cela signifie que tous les hommes ayant pris d_1 et souffrant de e_1 ont également pris d_2. On peut s'interroger sur les contributions respectives de d_1 seul, et de d_1 et d_2 pris ensemble en comparant leur PRR. On constate alors $PRR(d_1 e_1 M) \leq PRR(d_1 d_2 e_1 M)$ car, à sous-population constante, le PRR d'un non-fermé est toujours inférieur ou égal au PRR de son fermé, La prise concomitante de d_1 et d_2 contribue donc plus à l'apparition de e que d_1 seul.

Le critère de pertinence d'un signal ou d'une interaction est $PRR \geq 2$, $\chi^2 \geq 4$ et $support \geq 3$. Nous étendons la définition du PRR aux signaux et interactions contenant des attributs démographiques :

$$PRR(d, e, W) = \frac{P(e|dW)}{P(e|\overline{d}W)} \quad PRR(d_1 d_2, e, W) = \frac{P(e|d_1 d_2 W)}{P(e|\overline{d_1 d_2}W)}$$

5.2 Priorités entre signaux et interactions

Afin de limiter le nombre de signaux et interactions présentés à l'expert, nous établissons des règles de priorités entre signaux et interactions *comparables*. Cela concerne les signaux et interactions portant sur une même sous-population afin de ne pas comparer simultanément l'évolution de deux paramètres distincts.

Une interaction $(d_1 d_2 e W)$ est sélectionnée pour être présentée à l'expert, si elle est pertinente et prioritaire sur ses signaux comparables $(d_1 e W)$ et $(d_2 e W)$, i.e. si son PRR est supérieur aux PRR des deux signaux.

Un signal $(d_1 e W)$ est sélectionné pour être présenté à l'expert, s'il est pertinent et prioritaire sur au moins une de ses interactions comparables $(d_1 d_i e W)$, i.e. si son PRR est supérieur au PRR d'au moins une de ses interactions comparables.

6 Aide à l'évaluation et expérimentation

Nous montrons comment notre approche permet aux experts de « contextualiser » les signaux et interactions extraits en les replaçant dans leur contexte d'extraction au moyen de la représentation synthétique de la base que constitue le treillis. En complément des mesures de disproportionnalité, notre approche fournit des arguments aux experts leur

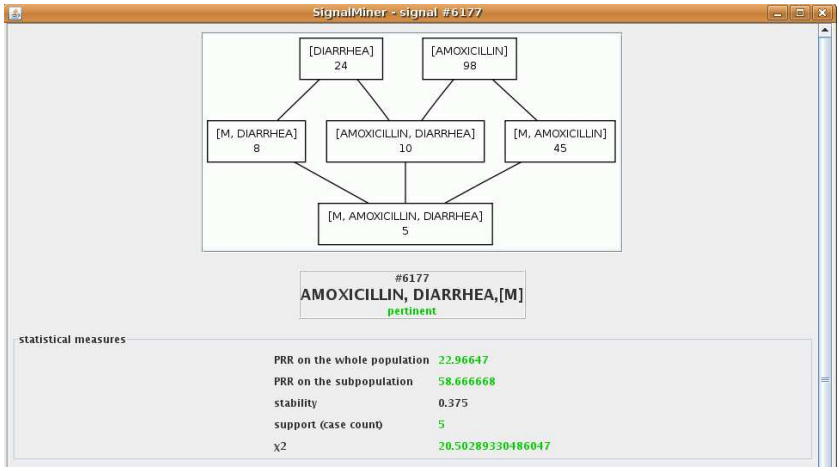

FIG. 2 – Interface utilisateur pour l'évaluation d'un signal

permettant de décider si un signal ou une interaction nécessite une expertise approfondie.

6.1 Visualisation et navigation

L'idée centrale est de considérer le treillis de concepts comme une représentation synthétique de la base. Une sous-partie du treillis relative au signal ou à l'interaction à évaluer est présentée à l'expert afin de resituer le signal ou l'interaction dans son contexte d'extraction.

La figure 2 représente l'interface présentée à l'expert pour l'évaluation du signal (d, e, W) où d est `amoxicilline`, e est `diarrhée` et W `homme`. Une sous-partie du treillis est affichée, où l'on peut voir le concept c_{deW} d'intension (deW) en bas, les concepts $\mu(d)$ et $\mu(e)$ en haut, et tous les concepts figurant sur les chemins allant de c_{deW} vers $\mu(d)$ et $\mu(e)$. Les concepts sont étiquetés par leur intension et leur support. Grâce à ce graphe, les experts peuvent observer la distribution des rapports dans la base : 24 patients souffrent de `diarrhée`, 98 ont pris de l'`amoxicilline`, 10 ont pris de l'`amoxicilline` et souffrent de `diarrhée` (signal (d, e)) dont 5 hommes (signal (d, e, M)). Les experts peuvent comparer les valeurs de PRR des signaux (d, e) (« PRR on the whole population ») et (d, e, M) (« PRR on the subpopulation ») et comprendre pourquoi, dans cet exemple, $PRR(d, e, M) \geq PRR(d, e)$. On observe que 5 hommes ont pris de l'`amoxicilline` parmi les 8 souffrant de `diarrhée`, i.e. la majorité d'entre eux. À l'inverse, seuls 10 patients ont pris de l'`amoxicilline` parmi les 24 souffrant de `diarrhée`. Ainsi, l'attribut démographique `M` accroît l'intensité du signal sur la sous-population des hommes. De plus, il est possible de comparer les valeurs de PRR sur différentes sous-populations (cf. Figure 3). Le même signal sur la population féminine montre un PRR plus faible (13,38). En effet, on a σ(`diarrhée`, `F`)=16 and σ(`amoxicilline`, `diarrhée`, `F`)=5. Le poids des patientes traitées à

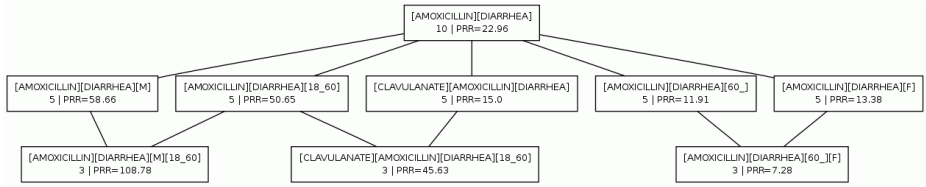

FIG. 3 – Évaluation sur différentes sous-populations

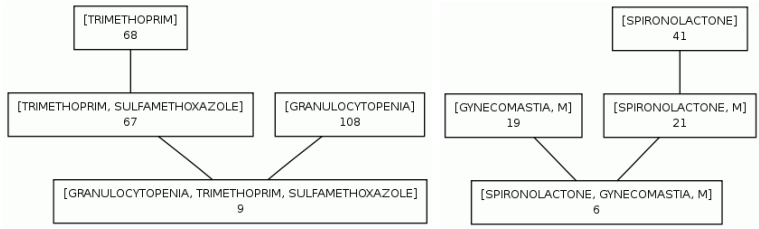

FIG. 4 – Exemple de données bruitées (gauche) et d'un effet indésirable masculin (droite)

l'`amoxicilline` parmi les femmes souffrant de `diarrhée` est plus faible que chez les hommes, et que dans la population totale.

Ainsi, la représentation synthétique que constitue le treillis permet aux experts de comprendre les raisons pour lesquelles un signal ou une interaction est plus prononcé sur une sous-population donnée. De plus, certaines situations comportent effets indésirables spécifiques à une sous-population, comme la `gynécomastie`, spécifique aux hommes (cf. Figure 4). Prendre en compte cette spécificité dans le calcul du PRR est nécessaire pour obtenir une mesure de disproportionnalité fiable.

6.2 Indices pour la détection du bruit

Le triméthoprime (d_1) et le sulfaméthoxazole (d_2) sont toujours administrés conjointement dans le traitement du Sida. Ainsi, un unique concept $\mu(d_1) = \mu(d_2)$ devrait figurer dans le treillis. Or on constate $\mu(d_2) \leq \mu(d_1)$ avec $\sigma(\mu(d_1)) = 68$ et $\sigma(\mu(d_2)) = 67$ (voir Figure 4). Cela signifie que, parmi les patients ayant pris d_1, un seul n'a pas pris également d_2, correspondant sans doute à un rapport erroné. L'indice de stabilité (Kuznetsov, 2007) d'un concept peut identifier ce type de situation. La stabilité quantifie la capacité d'un concept à demeurer persistant après la suppression d'objets de son extension. Ici, $\mu(d_1)$ a un faible indice de stabilité car la suppression du rapport erroné provoque la disparition de ce concept puisqu'alors d_1 n'est plus fermé et $d_1'' = \{d_1, d_2\}$. Ainsi $\mu(d_1)$ se confond alors avec $\mu(d_2)$. Un faible indice de stabilité pour un concept doit attirer l'attention des experts sur la possible présence de rapports bruités dans son extension.

TAB. 3 – Répartition des signaux extraits après leur évaluation par l'expert

catégorie	nb	
connu (documents de référence)	502 (89%)	vrais positifs
connu (sous une forme similaire)	24 (4%)	
effet à l'origine de la prescription	3 (1%)	faux positifs
dû à un médicament concomitant	9 (2%)	
inconnu	27 (4%)	expertise approfondie nécessaire

6.3 Expérimentation

Nous avons appliqué notre méthode à un extrait de la base de l'Hôpital Georges Pompidou. Cet extrait contient 3249 cas, 976 médicaments, 573 effets indésirables, et 5 attributs démographiques binarisés (deux pour le sexe et trois pour l'âge ($<$18, [18-60], $>$60). Le treillis correspondant contient 13178 concepts, dont 6788 ont un support d'au moins 3 rapports. Parmi les 2812 signaux candidats, 565 ont été sélectionnés comme signaux potentiels, et 102 interactions potentielles ont été retenues parmi 836 candidates. Notons qu'un espace de recherche exhaustif aurait engendré 976*573*(2+1)*(3+1) signaux candidats et ((976*975)/2)*573*(2+1)*(3+1) interactions candidates.

On observe que seuls 29% des signaux extraits concernent la population globale, le autres comportant un (49%) ou deux (22%) attributs démographiques. Leur validation *a posteriori* montre la pertinence de la prise en compte des aspects démographiques et l'efficacité de notre méthode. En effet, dans la majorité des cas, les attributs démographiques associés à un signal potentiel constituent effectivement un facteur aggravant connu. Ainsi, l'hypertension pulmonaire associée aux suppresseurs de l'appétit de la famille des amphétamines est observée chez les femmes de 18 à 60 ans.

Après évaluation, les signaux sont classés par l'expert en 5 catégories (cf. tableau 3). Les catégories 1 et 2 contiennent les vrais positifs, 3 et 4 les faux positifs et 5 les signaux potentiels inconnus, i.e. jamais signalés dans la littérature et nécessitant des expertises approfondies.

Les vrais positifs sont conformes aux résultats obtenus par les méthodes traditionnelles et aucun vrai positif n'est absent. Les faux positifs sont fréquents en détection du signal et certains d'entre eux sont bien connus. Ainsi le signal (`hydrochlorothiazide`, `toux`) est détecté car ce médicament et cet effet indésirable apparaissent souvent ensemble. Cependant, la toux est causée par les inhibiteurs de l'enzyme de conversion (IEC) pris concomitament avec l'hydrochlorotiazide. Du fait de la présence dans la base de plusieurs IEC différents d_i, chaque couple (d_i,`toux`) apparaît moins souvent que le couple (`hydrochlorothiazide`, `toux`). Ainsi, seul ce dernier signal est détecté. Une solution consiste à introduire des classes des médicaments, telles que la classe `IEC`, comme attributs supplémentaires avec oI`ACE` pour chaque rapport contenant un médicament de la famille des IEC. Des signaux de la forme (`IEC`, `toux`) pourraient alors être détectés.

7 Conclusion

Nous avons présenté une nouvelle méthode de détection automatique du signal mettant l'accent sur l'aide à l'évaluation par l'expert. L'espace de recherche peut être réduit aux motifs fermés, réduisant considérablement le nombre de signaux et interactions testés. Contrairement aux méthodes existantes, les aspects démographiques sont pris en compte afin d'identifier des sous-populations à risque.

Ce travail est un premier pas vers le développement de méthodes hybrides alliant approches numériques fondées sur les mesures de dispoportionnalité et approche symbolique fondée sur les treillis de concepts. Une expérimentation menée sur un extrait de la base de l'Hôpital Georges Pompidou montre qu'aucun signal attendu n'est manquant et que les résultats sont conformes à ceux des méthodes traditionnelles. Le treillis de concepts permet d'étoffer l'éventail d'arguments nécessaires aux experts pour évaluer si un signal ou une interaction nécessite une expertise approfondie. De plus, il offre le cadre permettant d'introduire une classification des médicaments afin de réduire le nombre de faux positifs, ce qui fait l'objet de nos prochaines recherches.

Références

BATE A., LINDQUIST M. & EDWARDS I. R. (2008). The application of knwoledge discovery in databases to post-marketing drug safety : example of the WHO database. *Fundamental & Clinical Pharmacology*, **22**(2), 127–140.

DUMOUCHEL W. (1999). Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, **53**(3), 177–190.

EVANS S. J. W., WALLER P. C. & DAVIS S. (2001). Use of proportional reporting ratios for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, **10**(6), 483–486.

GANTER B. & WILLE R. (1999). *Formal concept analysis : Mathematical Foundations*. Berlin : Springer.

HAUBEN M., MADIGAN D., GERRITS C. M., WALSH L. & PUIJENBROEK E. P. V. (2005). The role of data mining in pharmacovigilance. *Expert Opinion on Drug Safety*, **4**(5), 929–948.

KUZNETSOV S. (2007). On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, **49**(1-4), 101–115.

MEYBOOM R. H., EGBERTS A. C., EDWARDS I. R., HEKSTER Y. A., DE KONING F. H. P. & GRIBNAU F. W. J. (1997). Principles of signal detection in pharmacovigilance. *Drug Safety*, **16**(6), 335–365.

ROUX E., THIESSARD F., FOURRIER A., BÉGAUD B. & TUBERT-BITTER P. (2005). Evaluation of statistical association measures for the automatic signal detection generation in pharmacovigilance. *IEEE Transactions on Information Technology in Biomedicine*, **9**(4), 518–527.